13th ICCRTS
"C2 for Complex Endeavors"

Collaborative Learning Agent (CLA) for Trident Warrior

Submission 204

Topic 9: Collaborative Technologies

Ying Zhao, Chetan Kotak and Charles Zhou
Quantum Intelligence, Inc.
charles.zhou@quantumii.com

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **JUN 2008**	2. REPORT TYPE	3. DATES COVERED **00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE
Collaborative Learning Agent (CLA) for Trident Warrior

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Quantum Intelligence, Inc.,3375 Scott Blvd, Suite 100,Santa Clara,CA,95054

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
13th International Command and Control Research and Technology Symposia (ICCRTS 2008), 17-19 Jun 2008, Seattle, WA

14. ABSTRACT
Collaborative Learning Agent (CLA) is a technology selected for Navy on Trident Warrior ?08, which is an annual FORCEnet SEA Trial. The theme for ?08 is ?Maritime Domain Awareness?. The objective is to demonstrate a set of CLAs in a distributed network to learn behavior patterns from historical MDA data and then apply them for search, prediction, and identification of anomalies and reasons that might cause the anomalies, e.g. weather or potential terrorist activities. We will show collaborating with three MDA participants (Navy, Coast Guard and Police) using unstructured data sources as the bases for normal behavior profiles. A new real-time observation is compared with the normal profiles. An anomaly meter reports and shows if the new observation is an anomaly and why. The TW08 effective attributes include ?capable, accurate, usable and relevant? to evaluate CLA as follows: ? capable: agent learning and prediction from unstructured data ? accurate: compare with predictions with the ones from human analysts ? usable: easy-of-use in interface, visualization and display ? relevant: does CLA predict anomaly or interesting MDA behavior We will summarize the evaluation results in terms of these attributes.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

Abstract

Collaborative Learning Agent (CLA) is a technology selected for Navy on Trident Warrior ’08, which is an annual FORCEnet SEA Trial. The theme for ’08 is “Maritime Domain Awareness”. The objective is to demonstrate a set of CLAs in a distributed network to learn behavior patterns from historical MDA data and then apply them for search, prediction, and identification of anomalies and reasons that might cause the anomalies, e.g. weather or potential terrorist activities. We will show collaborating with three MDA participants (Navy, Coast Guard and Police) using unstructured data sources as the bases for normal behavior profiles. A new real-time observation is compared with the normal profiles. An anomaly meter reports and shows if the new observation is an anomaly and why. The TW08 effective attributes include “capable, accurate, usable and relevant” to evaluate CLA as follows:

- capable: agent learning and prediction from unstructured data
- accurate: compare with predictions with the ones from human analysts
- usable: easy-of-use in interface, visualization and display
- relevant: does CLA predict anomaly or interesting MDA behavior

We will summarize the evaluation results in terms of these attributes.

1. Objective

The goal of this paper is to employ a collaborative learning agent (CLA) in a context of Maritime Domain Awareness (MDA) in a Trident Warrior exercise to derive behavior patterns from historical MDA data, and use patterns in predictive analysis for anomaly detection. The specific questions to answer for this objective is

- Is the intelligent agent in CLA capable of learning from unstructured, historical information, for example, chat logs from all TW participants?
- Is CLA capable of prediction from unstructured data?
- Does CLA predict relevant anomalies or interesting MDA behavior?
- Is CLA accurate when its predictions are compared with predictions from human analysts?

2. Background

Port security is important. MDA is a critical component of the US national security strategies. MDA is defined as the effective understanding of anything associated with the global maritime domain that could impact the security, safety, economy, or environment of the United States. It is required to deploy the full range of its operational assets and capabilities to prevent the maritime domain from being used by terrorists, criminals, and hostile states to commit acts against the United States, its people, economy, property, territory, allies, and friends, while recognizing that maritime security policies are most effective when the strategic importance of international trade, economic cooperation, and the free flow of commerce are considered appropriately.

The critical business level needs of MDA include:

- Provide extended fusion and analytical capabilities through improved automation with broader crew and cargo coverage

- Provide historical trend analysis and behavior prediction that extends beyond vessel to nodal analysis
- Provide cross data type (vessel, cargo, crew, infrastructure) anomaly detection
- Provide improved threat recognition
- Provide expanded data sharing with inter-agency and coalition partners

Specifically, requirements are as follows:

- Fusion of multiple data sources across the maritime knowledge base, incorporating both current and archived data.
- Technologies that employ use of rule sets in decision making
- Improvements to current capability to identify, locate, track and target threats
- Algorithms that provide behavior prediction and pattern recognition applicable to vessel tracking, cargo
- Advanced data mining techniques

These requirements call for a total integration of machine learning, pattern recognition, information mining and search into a single collaboration technology for the total benefits of knowledge management for Maritime Domain Awareness.

One of drawbacks in current data mining is that tools are too complex and require professionals to run them. Mining results are not readily used in the search, which is a frequently used form of knowledge discovery and management. One drawback in current search engines is that the current search engines usually sort documents based on the popularity of documents, therefore they are not suitable to the applications that require looking for new, interesting and unique information which are not be popular or known by many people, for example, the anomaly detection application for MDA. Sorting information based on the degree of anomalousness has the potential to provide predictions, early warnings, and valuable business opportunities. The current search engines also require linked documents or databases for computing the relevance ranking which may not widely available for many domains except for the internet. They also require data to be copied in a centralized location, for example, current search engines usually crawl web pages into their servers and then index them for search. However, in real-life, the original data, especially unstructured data, are often generated and located in distributed organizations, servers and computers. Copying data into a centralized location is very expensive. For this reason, data warehousing projects are usually very expensive for organizations. Since original data sources are distributed, and organizations need to generate distributed indexes but share them globally as if it is a single index.

3. Collaborative Learning Agent (CLA)

A single Collaborative Learning Agent learns and discovers knowledge and behavior patterns from historical data and then applies the patterns for identification of patterns in the new data. The knowledge patterns, discovered automatically using machine learning and pattern recognition methods, include the following patterns

- Similarity patterns, i.e. group similar data
- Correlation patterns, i.e. find hidden relationships among data
- Predictive patterns, i.e. make predictions based on historical data

- Recommendation patterns, i.e. make predictions when little or no historical data is available.

A set of networked Collaborative Learning Agents (CLAs) forms an agent network include the following capabilities

- Text mining: extract concepts and meaning clusters based on contexts
- Machine meaning learning: extract knowledge patterns that link meaning to raw text or data observation
- Collaborative meaning search: incorporate human and machine a single loop to form a collaborative network to search and enhance the meaning iteratively.

A text mining technique, context-concept-cluster (ccc) model (US Patent Pending), is implemented in the CLAs. The advantage of such a text mining technique over the traditional information retrieval is to capture the cognitive level of understanding of text observations.

The process of machine learning meaning from human labeled data - supervised learning is illustrated in Figure 2. A train data set with both observations and their labeled meaning are presented to a machine learning system. The system first generates a text mining model which groups the data into categories by similarity. The system then generates a correlation model between the categories and meaning labeled by human analysts. The system also leaves out a held-out test data set for testing and evaluating, where the test set is fed into the same model and a meaning is predicted for each sentence in the test set. The predicted meaning is then compared with the real meaning labeled by human analysts to evaluate how accurate the meaning is predicted.

- Train data
  - Samples of the human labeled meaning of data
- Test data
  - Left out samples

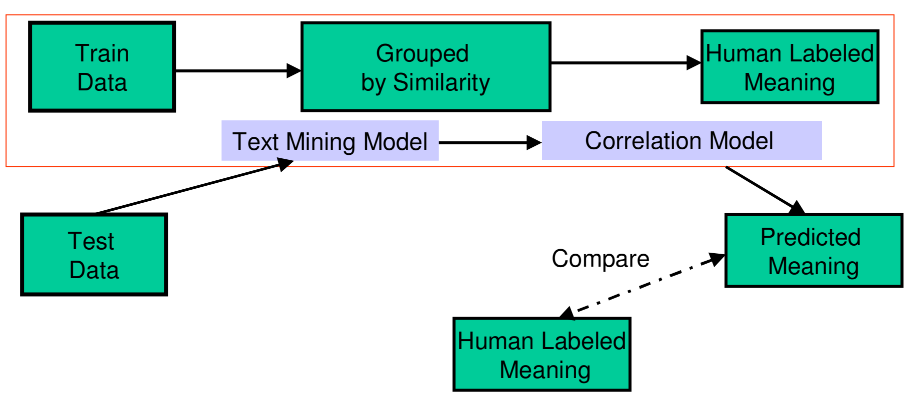

Figure 1: The process of supervised machine meaning learning

In real-life, however, human labeled meaning is expensive to obtain, therefore, it is more important to develop unsupervised learning to achieve the same goal. Here we want to show that CLAs are able to perform an unsupervised learning and categorize incoming information into four categories that are associated with the cognitive levels of human understanding and relevance as follows.

- Anomaly: showing an input is a relevant and interesting anomalous information
- Relevant: showing an input highly correlated to the agent networks' knowledge patterns.
- Low: between relevant and anomaly
- Totally anomaly – Nothing relevant: showing an input is an anomaly but not relevant.

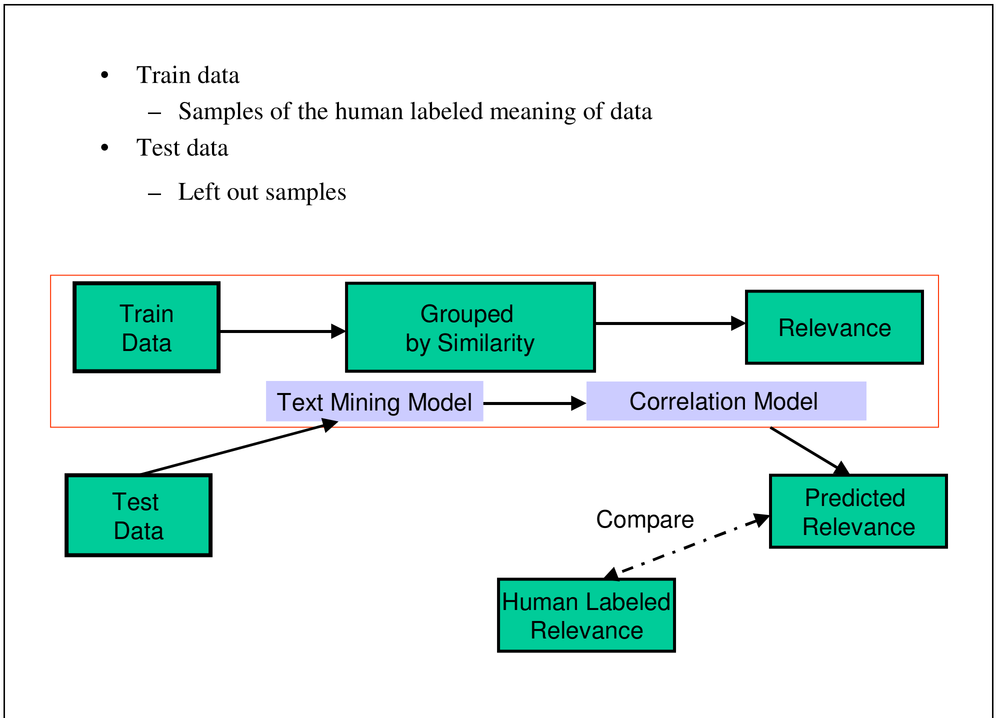

Figure 1: The process of unsupervised machine meaning learning

As shown in Figure 2, in the training part, human labeled meaning is not needed anymore: the CLAs automatically generate the relevance groups. Human labeled relevance is only used in the test data in order to validate if the predicted relevance from the CLA model is accurate.

A collaborative meaning search is further used to improve meaning prediction. Each agent (either human or machine) generates its own meaning model of assigning (predicting) a meaning to an input. Each agent also holds a peer list showing how an agent is socially connected. The true meaning of a piece of information is categorized with the combination of predictions from an agent's individual meaning model and its social network.

A CLA works with structured, unstructured data sources or combination of both. Structured data sources include excel, databases and XML data. Unstructured data sources to CLAs include free text (e.g. email, conversation transcripts and text chat), word, html, pdf, and ppt documents. In the context of MDA, port security and law enforcement, the data sources could be incident, trouble or suspicious activity reports from a social network, which are typically of paper or web forms that combine check boxes (structured) and free-text descriptive fields (unstructured) collected by field agents or reported by ordinary citizens.

A CLA network is an architecture for knowledge gathering, creation and dissemination network that allow collaborative users to integrate, analyze, find and understand information from a distributed network. The CLA architecture is different from current search engines in two ways: a) Indexes embedded in agents are distributed and customized to the data, learning and knowledge patterns of a local environment. This allows all data / index providers to maintain their own data locally in their distributed environment; b) Using semantic machine understanding makes it possible to search for new, interesting and unique information rather than popular information as in a current search engine, therefore, it is adaptive, semantic, distributed and collaborative.

4. Applications to MDA

We use the CLA technology to the MDA by deploying a set of networked CLAs. Specifically, CLA improves fusion/correlation across maritime knowledge bases. CLA uses advanced data mining techniques to discover the patterns and rules for decision and sense making. CLA will improve current capability to identify, locate and target threats, and provide behavior prediction and pattern recognition applicable to vessel tracking, cargo.

For example, to prevent potential threats and events of terrorism related to the global maritime domain, DOD, federal, state and local agencies have to work together to make situation awareness effectively based on a Common Operational Picture (COP). The unclassified COP integrates dynamic and static data regarding a variety of coordination efforts for boarder and force protection. Some of the COP tools already provide coherent pictures through geospatial data and visualization. Meanwhile, these tools provide dynamic near-real-time data feeds, for example, chats and trouble calls from the collaborators as unstructured data. This data currently is being manually analyzed by human analysts. Incorporating the CLA, knowledge patterns are learned from historical data analyses and then applied to new data to understand intent and recommend what action to take. Therefore, an agent can be trained to reduce manpower for common decision/sense making tasks and free people up for other tasks.

MDA requires complex physical and social networks that involve many human interfaces and machines. In real-time there are numerous trouble calls. These trouble calls need to be assessed, analyzed and resolved automatically before their operational impact takes place. Currently, trouble calls have to be analyzed by human analysts. Knowledge patterns can be learned and transferred to computer agents. Therefore manpower required for handling the trouble calls can be greatly reduced.

In summary, CLA can be used to analyze and mine information (structured and unstructured) collected in the vessel tracking, cargo monitoring, and people screening/identity management, predict threats and anomalies. CLA can also provide capabilities for collaboration and visualization for MDA operations. CLA can also address inter-agency and coalition interoperability and information sharing. Data sources such as NCIS (Naval Crime Investigation Service) containing good information for MDA. ESP Portal is a new collaboration tool for international border/force protection, where international participants such as Australian, Japanese or Singapore border controls or coast guards can all join the portal to provide data. CLA is designed to leverage the collaborations. CLA's architecture also addresses the integration of the future US NAVY CANES architecture. Specifically, for example, in Trident Warrior 08, the plan is to obtain unclassified, unstructured data such as chat logs and event trace data exported from the tools installed in CANES through web services.

5. CLA Architecture for Trident Warrior

In a top level, the objectives of using CLAs for TW is to first identify and obtain data samples from three MDA partners, for example, participants representing inter-agency collaborations among navy, police, coast guard. Then we apply a single CLA to each MDA data source to discover knowledge patterns representing normal behavior patterns. The discovered knowledge patterns from historical data are served as the normal profiles for new data to compare with. The knowledge patterns are stored within a search index in a search network including multiple indexes from multiple CLAs. When a piece of real-time information is newly observed, it goes through the search network, the network returns a report of search results, showing if the new information is correlated with the normal behavior profiles and what degree of the correlation is, or if it is a totally new information or an anomaly. The process is illustrated in Figure 3.

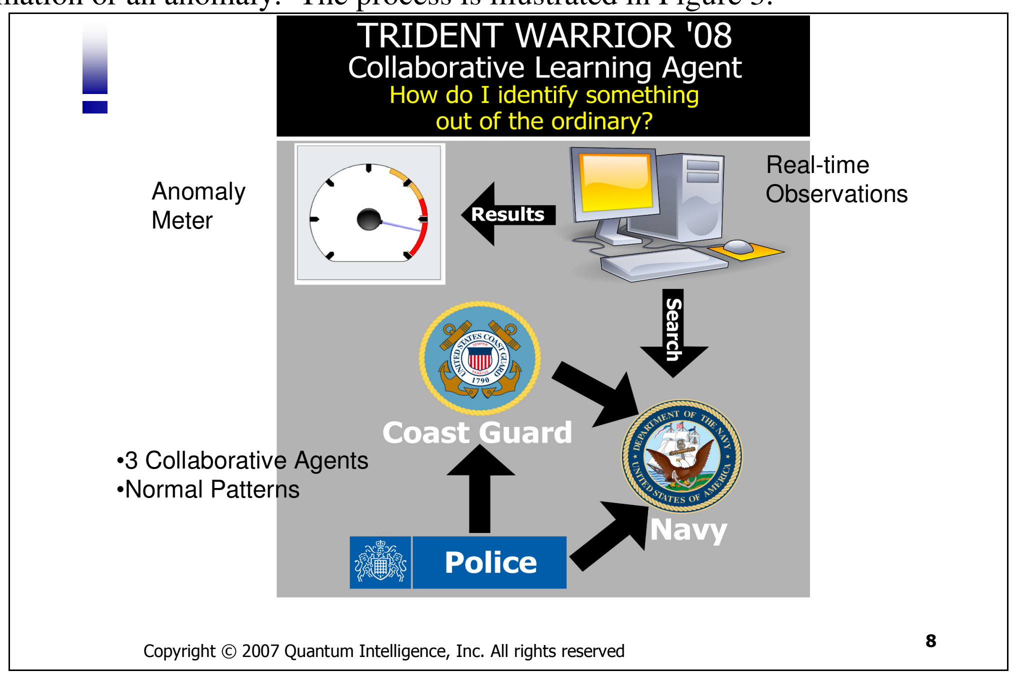

Figure 3: Trident Warrior Architecture

The three collaborative agents are installed in the Naval Postgraduate School (NPS) Monterey, California in the non-classified Internet to access the chat data representing three collaborating partners as follows:

- http://cla1.quantumii.com/NAVY/
- http://cla2.quantumii.com/USCG/
- http://cla3.quantumii.com/CIVIL/

6. Data

Since we are still in the process of TW08, we are using the TW07 chat data as an example to illustrate the process. There were four days of MDA exercise chat data: March 15, March 16, March 19, and March 20, 2007. We grouped the collaborators under the larger umbrella of local civil authorities, navy and coast guard as follows:
NAVY: NCWS(Naval Coastal Warfare Squadron), CNRMA(Commander, Navy Region Mid-Atlantic), ROC (Regional Operations Center),LCPO (Lead Chief Petty Officer), USNR (United States Naval Reserve), JTTF (Joint Terrorism Task Force), and NCIS (Naval Criminal Investigative Service)
USCG: USCG (United States Coast Guard) and PWCS (Ports, Waterways, Coastal Security)
CIVIL: Norfolk EOC (the City of Norfolk Emergency Operations Center) which includes both NPD (Norfolk Police Department) and NFD (Norfolk Police and Fire Departments). VB EOC (the City of Virginia Beach Emergency Operations Center which includes both Virginia Beach Police Department (regular and Marine units) and Virginia Beach Fire Departments, CBBT Police (Chesapeake Bay Bridge Tunnel Police), CBP(Customs & Border Protection), NRC (Nuclear Regulatory Commission),COTP (Captain of the Port)

The training model is built on the first three days (March 15, March 16 and March 19) and the last day (March 20) is used for test and validation.

•Training
–March 15, 16 and 19
•Test
–March 20

7. Results and Evaluation

Using CLAs, we were able to successfully discover two clusters of in the test set (March 20) that were not labeled initially by any human analysts as follow:

- TW scenario related, for example, a civil authority reports a suspicious activity, then a coast guard issues a security zone and a navy participant provides security escort for investigation
- TW technology related, for example, chat about test, problems and where to find maps etc.

Overall, we found the category “relevant” mostly belongs to the TW scenario and the category “low” mostly belongs to the TW technology. We did not find any interesting anomalies in the data and found some non-relevant anomalies, indicating most TW07 chat stayed in its objective of executing a scenario and evaluate the technologies.

The following metrics will be used in the evaluation.

7.1.1 Capable

Capability questions are

- Are anomalies being detected?
- Is anomaly meter showing any activity?

This will be a dual assessment consisting of both a gross measure and a more detailed measure. The gross measure will be real-time and will consist of counts per time period by the system user of anomalies detected (e.g., count during a 5-minute period of time every hour). The more detailed measure will be post-scenario analysis of the number of correct detections per total events (i.e., chat messages).

7.1.2 Accurate

Accuracy is as an analyst assessment of events counted as anomalies (% correct vs. false positives and false negatives). For example, the total test chats used is about *c*, *a* of them are labeled correctly by CLAs.

$$\text{Accuracy} = \text{\# of correct /total test chats} = a/c$$

7.1.3 Relevant

Relevancy is defined as how a human analyst interpretation of relevancy reported.
For example, if the total number of TW scenario sentences is *s* (labeled by a human analyst) and the number that are labeled correctly by CLAs is *r*.

$$\text{Relevancy} = r/s$$

7.1.4 Usable

Usability is as an analyst assessment of clarity of display, extent to which trusted, ease of accessing the detailed data

8. Conclusion

We showed proof-of-concept for the CLA technology as a set of networked agents learning behavior patterns from historical MDA data and then applying them for prediction and identification of anomalies and reasons that might cause the anomalies for new data.

9. Acknowledgements

Thanks to SBIR ONR Contract No. N00014-07-M-0071. Thanks to Dr. Shelley Gallup at Naval Postgraduate School and Dr. Nancy Heacox at Pacific Science & Engineering Group, Inc. for providing TW07 data and valuable discussions.